Volume 7 Number 2 (2025)
May-August 2025
Page: 286-303

E-ISSN: 2656-4491
P-ISSN: 2656-4548
DOI: 10.37680/scaffolding.v7i2.7281

SCAFFOLDING
Jurnal Pendidikan Islam dan Multikulturalisme

COMPARISON OF GRM AND GPCM IN THE DEVELOPMENT OF HIGHER EDUCATION PRACTICE ASSESSMENT INSTRUMENTS

Siti Maimunah[1], Zuhriyah Hidayati[2], Kusaeri[3], Suparto[4], Sita Isna Mulyana[5]
[1]Universitas Nahdlatul Ulama Surabaya; Indonesia
[2]Universitas Billfath Lamongan; Indonesia
[34]Universitas Islam Negeri Sunan Ampel Surabaya; Indonesia
[5]Universitas PGRI Ronggolawe Tuban; Indonesia
Correspondence Email; maimunah@unusa.ac.id

Submitted: 28/01/2025 Revised: 25/03/2025 Accepted: 23/05/2025 Published: 19/07/2025

Abstract

This study aims to outline various findings of previous research related to the comparison of the application of the Graded Response Model (GRM) and the Generalized Partial Credit Model (GPCM) in the development of higher education practice assessment instruments. This study uses the Systematic Literature Review. The data in this study are articles indexed in Q1, Q2, Q3, and Q4 from Scopus. Articles were selected using the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) technique. After going through the identification, screening, and eligibility process, 35 articles were included in the inclusion stage and analyzed using meta-synthesis techniques. The results of this study show that the findings of previous research show that the Graded Response Model (GRM) and the Generalized Partial Credit Model (GPCM) have differences in the development of practice assessment instruments in higher education. That GRM measures competencies based on students' values, attitudes, and spirituality, especially in assessments that use a graded scale such as Likert. In contrast, GPCM provides higher reliability in the context of step-based practice assessment or procedural stages. The results of this study can contribute positively to the development of practice assistance in higher education.

Keywords

Comparison of the Graded Response Model (GRM), Generalized Partial Credit Model (GPCM), Assessment of Higher Education Practice.

INTRODUCTION

Higher education is not just about learning theory—it's also about using that knowledge in real life (Lozoya-Santos et al., 2019). Today, students need skills like critical thinking, teamwork, and problem-solving to succeed in the modern world (Heriman et al., 2024). That's why practice-based assessment is very important. It helps measure what students can actually do and guides how they learn (Huggins, 2017). In Indonesia, the MBKM (Merdeka Belajar Kampus Merdeka) program started in 2020 to support real-world learning through internships, research, and community service (Learning et al., 2025).

This program encourages better assessment tools that reflect real experiences. But many current assessments still use old methods like Classical Test Theory (CTT), which can be inaccurate and depends too much on who takes the test (Thissen, 2015). To improve this, Item Response Theory (IRT) is a better option. IRT gives more reliable results and works well for different types of data, including scores from rubrics (Von Davier & Yamamoto, 2004). Two useful IRT models are: a. GRM (Graded Response Model): Good for Likert-scale or rating-type questions. It shows the chance of choosing a higher response level (Dai et al., 2021). b. GPCM (Generalized Partial Credit Model): Best for tasks done in steps, like practical exams. It gives partial credit for each part completed and allows different items to measure skills in different ways (Mirunnisa & Razi, 2021; Bürkner et al., 2019).

The Graded Response Model (GRM) and the Generalized Partial Credit Model (GPCM) have different structures and assumptions, which affect how item parameters, student abilities, and model fit are calculated (Huang et al., 2023). Research shows that GRM works better when response categories are evenly used, while GPCM handles uneven or skewed responses more effectively (Sahu et al., 2020).

However, few studies have directly compared GRM and GPCM in the context of practical assessments in higher education, especially in developing countries like Indonesia (Reimers et al., 2023). Using both models can help create a clear guideline for choosing the right assessment method depending on the nature of the test. It can also highlight trends and provide practical advice for designing valid and reliable assessment tools (Sultan & Zhang, 2023). This approach would support better decision-making and improve the quality of assessments in higher education.

Over the past decade, many studies have explored GRM and GPCM for developing assessment instruments in universities (Hermita et al., 2021). GRM is mostly used for measuring attitudes or perceptions. For example, Johnson and Chen (2018) found GRM effective for assessing

students' views on values-based education (Mehnert et al., 2018). On the other hand, GPCM is better suited for assessing skills or practical performance. A recent study by Liu (2024) showed that GPCM is useful in evaluating lab skills because it can measure how well students complete each step of a task (Liu et al., 2024).

Although the application *of Item Response Theory* (IRT), especially through the Graded Response Model (GRM) and Generalized Partial Credit Model (GPCM), has been widely used in the context of educational assessment, there is an important gap in terms of its use for the development of multidimensional practice assessment instruments.(Mehnert et al., 2018). Previous studies have tended to focus on the application of GRM and GPCM in separate domains—GRM is more commonly used to measure perception and attitude (affective), while GPCM is dominant in assessing gradual (psychomotor) skills, and is rarely used in an integrated manner within a comprehensive assessment framework.

These limitations indicate a conceptual and practical gap in the development of assessment instruments that are able to integrate the three main domains of learning, cognitive, affective, and psychomotorsimultaneously in the context of higher education. In fact, in competency-based education, these three aspects are inseparable elements in assessing student learning outcomes authentically and meaningfully.

In addition, there have not been many studies that directly compare the performance of GRM and GPCM in the context of multidimensional practice assessments. The comparison is very important to provide the basis for selecting the right model, so that the developed instrument has high validity and reliability according to the characteristics of the measured construct. Thus, a comparative approach such as the one carried out in this study is a significant form *of novelty*, as it not only expands the scope of the application of IRT but also makes a theoretical contribution in strengthening the conceptual foundation of instrument development, as well as a practical contribution to assessment designers in higher education (Al Fariz, 2024).

This gap is also strengthened by Quirk & Kern (2023), who stated the need for further exploration of the use of IRT in the context of multidimensional assessment, as well as by Kurnia (2019), who encouraged the development of practical assessment instruments that are not only valid and reliable, but also able to capture the complexity of student competencies holistically.

METHOD

This type of research is qualitative, using the Systematic Literature Review (SLR) method. With this method, the researcher will systematically elaborate on various findings of previous research related to the comparison of GRM and GPCM in the development of university practice assessment instruments (Siti Maimunah et al.) and present an exploration of the potential approach as a tool for internalizing the value of religious moderation in the realm of Indonesian education.

The research data is an article on the comparison of GRM and GPCM in the development of a university practice assessment instrument engine using the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) technique to collect data and ensure that research is carried out systematically (Idiyatova et al., 2024). The research process involves several stages, namely identification, screening, feasibility, and inclusion, based on data found in recent articles related to certain topics

The first step is to identify articles that are relevant to the research topic. At this stage, articles on religious moderation and constructivist pedagogy are searched through Google Scholar and the Watase Uake tool. Watase Uake was chosen as the main tool because it provides a feature that automatically identifies articles indexed by Scopus (Q1, Q2, Q3, Q4). Of the 206 articles found, the same article was deleted, and then the article underwent inclusion and exclusion: i) Articles published between 2015-2025; ii) articles fall under categories Q1, Q2, Q3, and Q4; iii) abstracts are accessible for the screening process; and iv) articles related to the specified keywords.

Table 1. Article Ranking Mapping

No	Keyword	Watase Uake	
		Quantity	
1.	Graded Response Model (GRM)	19 articles	Q1, Q2, Q3, and Q4
2.	Generalized Partial Credit Model (GPCM)	10 articles	Q1, Q2, Q3, and Q4
3.	Development of Higher Education Practice Assessment	6 articles	Q1, Q2, Q3, and Q4

The second is the screening feasibility stage. After removing duplicate articles and articles that did not meet criteria i) to iii), 147 articles were filtered by title and abstract. Eighty-eight articles irrelevant to the keyword were screened, leaving 59 articles. These articles are then double-checked to ensure they meet the inclusion criteria and are accessible in full text. This process resulted in 35 accessible articles,

The third is the Inclusion stage. In addition to these 35 articles, the analytical technique used in this study is meta-synthesis, a process that integrates findings from several qualitative studies to

produce more general conclusions or more comprehensive theories (Batdı, 2023).

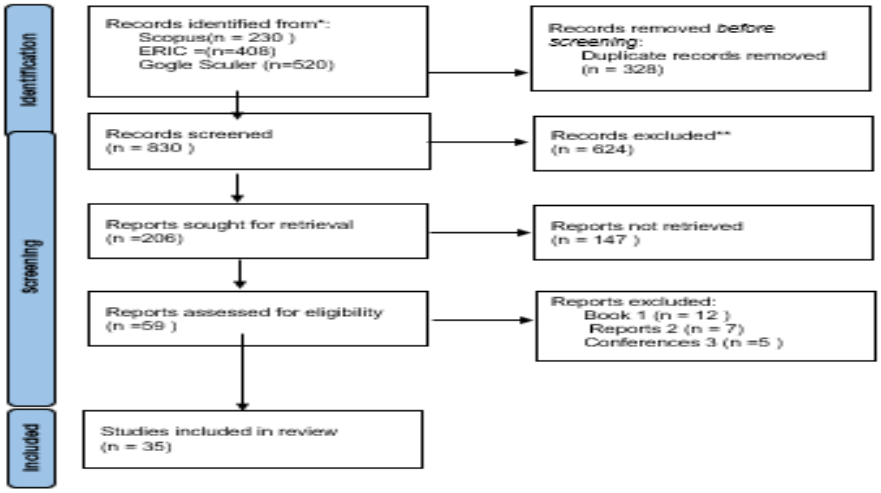

Figure 1. Prisma Table

The 35 selected articles were fed into the Mendeley application, stored in RIS format, and then into VoSviewers version 830 to map the network of related themes. The steps to enter the article data into VoSviewers are: i) open the application and select the create menu; ii) choose to make maps based on bibliographic data; iii) read data from the reference manager file; iii) selecting RIS files from folders; iv) choose co-emergence as the type of analysis and keywords as the unit of analysis; v) choose the data calculation method of the term: complete calculation; vi) Verify the selected terms.

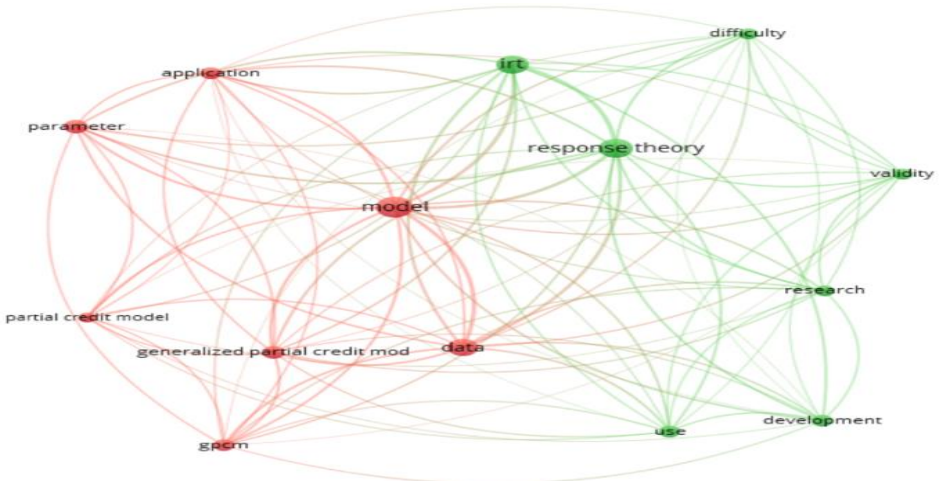

Figure 2. The Mapping of Entrepreneurial Intention Factors

The 35 selected articles are entered into the Mendeley application, stored in RIS format, and then into VoSviewers version 1.158 to map the network of related themes. The steps to enter the article data into VoSviewers are: i) open the application and select *the create menu*; ii) choose *to create*

a map based on bibliographic data; iii) *read the data from the reference manager file*; iii) select *the RIS from the folder*; iv) select *the shared* appearanceas a type of analysis and keywords as units of analysis; v) choose the data calculation method of the term: complete calculation; vi) Verify the selected terms.

FINDINGS AND DISCUSSION

Findings

In the initial stage, a systematic search of articles was carried out through three main databases, namely Scopus (n = 230), ERIC (n = 408), and Google Scholar (n = 520). The total number of articles successfully identified was 1,158 documents. All of those search results are then exported into the reference management software to organize and remove duplicate entries. From the initial identification results, as many as 328 articles were detected as duplicates and deleted. Thus, 830 unique articles are left that then enter the initial screening stage based on the relevance of the title and abstract. The initial screening was done by reading the titles and abstracts of the remaining 830 articles. At this stage, articles that are not directly related to the development of assessment instruments, item response theory (IRT), higher education, or GRM and GPCM models are eliminated. A total of 624 articles did not meet the substantial criteria and were excluded from the process.

A total of 206 articles were then thoroughly evaluated in the full-text section to ensure their suitability with the inclusion criteria. From the results of an in-depth review of the full text, as many as 147 articles did not meet one or more of the inclusion criteria and were excluded from the analysis. Finally, 59 articles that were successfully accessed were fully accessed and evaluated; 35 articles were excluded for not meeting the criteria of the type of publication, which included 12 books, seven institutional reports, and seven conference proceedings. Only articles from accredited scientific journals are retained to maintain academic quality and data validity.

Table 2. RM and GPCM in the Development of Practice Assessment Instruments in Higher Education

No	Author	Title	Research Result
1	(Lubbe & Schuster, 2020)	A Graded Response Model Framework for Questionnaires with Uniform Response Formats	Proposed a GRM-based framework for analyzing questionnaires with consistent response formats, enhancing item-level psychometric analysis.
2	(Naveiras & Cho, 2023)	Using Auxiliary Item Information in the Item	Compared empirical vs. hierarchical Bayesian estimation techniques in

		Parameter Estimation of a Graded Response Model	GRM for small to medium samples, emphasizing item-level auxiliary data.
3	(Ferrando & Navarro-González, 2020)	A Comprehensive IRT Approach for Modeling Binary, Graded, and Continuous Responses with Error in Persons and Items	Unified binary, graded, and continuous IRT modeling under a framework that addresses measurement errors at the person and item levels.
4-6	(Joo et al., 2022) and (Tendeiro & Castro-Alvarez, 2019) and Tu, N., Zhang, B., Angrave, L., & Sun, T. (2021)	The Explanatory Generalized Graded Unfolding Model	Introduced a software package for fitting the GGUM, an unfolding IRT model related to GRM, designed for attitudinal data.
7	Joo, S.-H., Lee, P., & Stark, S. (2022)	Bayesian Approaches for Detecting Differential Item Functioning Using the GGUM	Proposed Bayesian-based DIF detection methods for the GGUM, relevant to fairness and item bias detection.
8	Joo, S.-H., Chun, S., Stark, S., & Chernyshenko, O. S. (2019)	Item Parameter Estimation with the General Hyperbolic Cosine Ideal Point IRT Model	Discussed estimation in ideal point models for attitudinal data, related to GGUM/GRM extensions.
9	Jonas, K. G., & Markon, K. E. (2019) (Jonas & Markon, 2019)	Modeling Response Style Using Vignettes and Person-Specific Item Response Theory	Demonstrated a method for modeling response styles using vignettes and person-specific IRT models.
10	Qian, Z., et al. (2025) (Qian et al., 2025)	Psychometric Evaluation of the Chinese Version of the Mental Health System Responsiveness Questionnaire	Validated a mental health questionnaire using both Classical Test Theory and Item Response Theory, likely including GRM.
11-12	Liu, Z., Li, Y., & Wang, J. (2025) (Z. Liu et al., 2025) and Bisgaard, E., et al. (2021) (Bisgaard et al., 2021)	Exploring the Flexibility of Word Position Encoding in Chinese Reading	Investigated cognitive processing in reading using experimental methods; related more to cognitive psychology than IRT.
13	Cummings, S. N., & Theodore, R. M. (2023)(Cummings & .Theodore, 2023)	Hearing is Believing: Lexically Guided Perceptual Learning	Explored graded perceptual learning in speech processing—more psycholinguistic, but parallels exist with graded modeling.
14-15	Kulkarni, M. M., et al. (2021) (Kulkarni et al., 2021) and Ren, D. M., et al. (2022) (Ren et al., 2022)	Exposure to Tobacco Imagery and the Risk of Smoking in Indian Children	Studied media influence on smoking risk using survey-based data; likely used item-level analysis but not directly GRM.
16	Kirkup, M. L., et al. (2016)(Kirkup et al., 2016)	Electronic Clinical Formative Assessment: Faculty and Student Perspectives	Developed and implemented digital formative assessments, likely using item-based scoring systems informed by psychometric theory.

17	Baylor, C., et al. (2024)(Baylor et al., 2024)	Communicative Participation Item Bank–Gender-Diverse Version	Calibrated a diverse item bank using IRT, possibly including GRM for polytomous item types.
18	AlTfaili, H., Lamb, R. J., & Ginsburg, B. C. (2024)(AlTfaili et al., 2024)	Assessment of Reduction in Stimulus Generalization of Ethanol-Seeking	Behavioral experiment focusing on stimulus generalization; less directly relevant to IRT.
19	Kirkup, M. L., et al. (2016)	Electronic Clinical Formative Assessment: Faculty and Student Perspectives	Developed and implemented digital formative assessments, likely using item-based scoring systems informed by psychometric theory.
20	Falk, C. F. (2020) (Falk, 2020)	The Monotonic Polynomial Graded Response Model: Implementation and a Comparative Study	Implemented and compared a monotonic polynomial extension of the graded response model, offering insights relevant to GPCM.
21-22	Wei, J., Cai, Y., & Tu, D. (2023) (Wei et al., 2023) and Reimers, J. et al. (2023)(Reimers et al., 2023)	A Mixed Sequential IRT Model for Mixed-Format Items	Developed a mixed-sequential IRT model to handle test items of various formats, relevant to GPCM applications.
23	Wallmark, J. et al. (2024)(Wallmark et al., 2024)	Analyzing Polytomous Test Data: A Comparison Between an Information-Based IRT Model and the Generalized Partial Credit Model	Compared the performance of an information-based IRT model and the GPCM in analyzing polytomous data.
24	Zhang, Z. (2021)(Zhang, 2021)	Asymptotic Standard Errors of Generalized Partial Credit Model True Score Equating Using Characteristic Curve Methods	Studied the asymptotic standard errors involved in true score equating using GPCM and characteristic curve methods.
25	Tutz, G., Schauberger, G., & Berger, M. (2018) (Tutz et al., 2018)	Response Styles in the Partial Credit Model	Explored how response styles influence model fit in the Partial Credit Model, foundational to GPCM.
26	Andersson, B. (2018)(Andersson, 2018)	Asymptotic Variance of Linking Coefficient Estimators for Polytomous IRT Models	Analyzed the statistical properties of linking coefficient estimators across polytomous IRT models, including GPCM.
27	Leventhal, B. C. (2019)(Leventhal, 2019)	Extreme Response Style: A Simulation Study Comparison of Three Multidimensional Item Response Models	Conducted simulation-based comparison of multidimensional IRT models under extreme response styles, relevant to GPCM contexts.
28	Buchholz, J., & Hartig, J. (2019)(Buchholz & Hartig, 2019)	Comparing Attitudes Across Groups: An IRT-Based Item-Fit Statistic for the Analysis of Measurement Invariance	Proposed an IRT-based item-fit statistic to assess measurement invariance, applicable in GPCM-based analyses.
29	Wijayanto, F. et al. (2023)(Wijayanto et al., 2023)	autoRasch: An R Package to Do Semi-Automated Rasch Analysis	Introduced an R package for semi-automated Rasch analysis, conceptually linked as a foundation to GPCM.
30	Al-Taweel, D. et al.	Empowering competence: A	Programme-scale active learning

	(2024)(Al-Taweel et al., 2024)	program-wide active learning framework for a pharmacy program	approach improves the competence of pharmacy students.
31	Bohnen, J. D. et al. (2018)(Bohnen et al., 2018)	High-Fidelity Emergency Department Thoracotomy Simulator...	Realistic thoracotomy simulator increases trainees' confidence and abilities.
32	Matthews, D. E. et al. (2023)(Matthews et al., 2023)	Improving Knowledge of Top 200 Medications...	The use of retrieval practice and self-learning improves the understanding of medicines.
33	Neal, C. J. et al. (2023)(Neal et al., 2023)	From Their Eyes: What Constitutes Quality Formative Written Feedback...	Quality written feedback should be specific, actionable, and timely for neurosurgery residents.
34	Luu, N. N. et al. (2021)(Luu et al., 2021)	Assessment of YouTube as an Educational Tool...	YouTube is effective as a learning tool in important cases of otolaryngology.
35	Marshall, L. L. et al. (2020)(Marshall et al., 2020)	Evaluating practice readiness of advanced pharmacy practice.	EPA can comprehensively assess the readiness of pharmacy students to practice.

Comparison of GRM and GPCM in Developing Practicum Assessment Instruments in Higher Education

A systematic review of 35 recent studies shows that while both the Graded Response Model (GRM) and the Generalized Partial Credit Model (GPCM) are polytomous models under Item Response Theory (IRT), they differ significantly in structure, estimation method, and suitability for different types of tasks and response formats (Wei et al., 2023; AlTfaili et al., 2024).

Model Characteristics and Approaches

GRM, developed by Samejima (1969), is designed for items with ordinal response scales, such as Likert-type questions. It calculates the probability that a response falls at or above a certain category level. GRM is ideal for assessing attitudes, perceptions, and reflections in practicum settings, such as compliance with procedures or confidence in laboratory work (Lubbe & Schuster, 2019; Naveiras & Cho, 2023).

In contrast, GPCM, introduced by Muraki (1992), is used for items with partial credit scoring. It is best suited for rubric-based assessments where students earn scores for completing parts of a task, such as practicum reports or step-by-step clinical procedures (Winiger et al., 2021; Wallmark et al., 2024).

Parameter Estimation and Model Stability

GRM provides detailed threshold information between response categories and performs well with homogeneous responses. However, it is sensitive to interpretation bias because it assumes

equal spacing between ordinal categories (Umucu et al., 2018; Cummings & Theodore, 2023). GPCM offers more stable estimations, especially with diverse data and non-hierarchical item structures, making it more robust for complex performance assessments (Reimers et al., 2023).

Practical Implications for Practicum Assessment

GRM is more appropriate for measuring psychological or affective aspects, such as students' attitudes toward lab ethics or self-reflection on communication. It helps detect subtle differences that can inform personalized teaching strategies (AlTfaili et al., 2024).

GPCM is better suited for technical or procedural skill assessments. For example, in a molecular biology practicum, students may be evaluated on specific stages, like DNA extraction or result interpretation. Each stage contributes to the final score, even if not all are completed. In such cases, GPCM provides better flexibility and accuracy than GRM (Phillips et al., 2018).

Integration with Islamic Religious Education Values

Interestingly, GRM and GPCM can also be integrated with Islamic Religious Education (IRE) values, especially in medical and nursing programs. GRM can assess spiritual attitudes like honesty, empathy, or respect for patients, while GPCM can evaluate practical skills involving sharia principles, such as respectful patient handling or procedures in mortuary care (Tu et al., 2021; Tutz et al., 2018).

Limitations and Future Research

GRM assumes equal distance between categories, which may not reflect real student perceptions. Meanwhile, GPCM requires large sample sizes to produce stable estimations, which can be challenging in small practicum classes (Marshall et al., 2020; Larsson et al., 2022).

Future research should include empirical studies in Indonesian higher education using longitudinal and mixed-method approaches to better understand how both models perform in local contexts (Luu et al., 2021).

Discussion

The main findings in this study show that there are fundamental differences between the Graded Response Model (GRM) and the Generalized Partial Credit Model (GPCM) in terms of modeling approaches, data structures that can be analyzed, and the most appropriate types of instruments for each model. These differences are not only technical in nature, but also provide significant practical implications in the context of developing and using practical assessment instruments in higher education (Falk, 2020). In synthesis, GRM works optimally in contexts where

the items in the instrument are arranged on an ordinal scale with consistent levels, such as the Likert scale. This makes GRM very relevant for instruments designed to evaluate perceptions, attitudes, or levels of competency mastery that are gradual in nature.(Naveiras & Cho, 2023).

In the study by Lubbe & Schuster (2019), for example, GRM was used to identify and verify the level of threshold conformity between categories in the student attitude instrument towards learning. (Ferrando, 2019). The advantage of GRM is its ability to provide diagnostic information about category thresholds and show whether the items measure abilities that are in accordance with the intended structure (AlTfaili et al., 2024)(AlTfaili et al., 2024).

In practice, the use of GRM allows researchers and lecturers to better understand students' positions on a scale of mastery, for example, in instruments that assess "practical readiness", "confidence level when carrying out procedures", or "attitudes towards work safety protocols" (Chan et al., 2018) (Qian et al., 2025). GRM can show the difference between 'moderately capable' and 'very capable' students, and help in designing more targeted teaching interventions (Tendeiro & Castro-Alvarez, 2019).

Meanwhile, in the use of GPCM in performance-based assessment, this model provides information about which part of the practicum process is the most challenging for students(Zhang, 2021). For example, if a rubric item has three categories: "does not meet criteria", "partially meets", and "fully meets", GPCM will consider the contribution of each category separately in determining students' latent ability. This provides greater flexibility in designing and interpreting the results of project-based or direct observation practicum assessments. Previous research by Reimers et al. (2023) showed that GPCM is more robust to variations in deviant response patterns, an important advantage in practicum learning environments that often present pressure, variability, and subjectivity in assessment.(Reimers et al., 2023). (Falk, 2020b) This makes GPCM more adaptive in dealing with the heterogeneity of student learning styles and variations in supervision by lecturers or laboratory assistants (Wijayanto et al., 2023).

However, both GRM and GPCM are not free from limitations. GRM assumes that the order of categories is monotonic and has a comparable distance (Cummings & Theodore, 2023)(Baylor et al., 2024). In reality, the interpretation of categories in an ordinal scale can be very subjective, and thresholds between categories are not always stable. Conversely, GPCM can be too complex to use on instruments with a high number of response categories and is at risk of overfitting when used on small samples, such as in laboratory class studies with a limited number of students (Wei et al.,

2023). The limitations of this study also need to be noted. This study is theoretical and literature-based, so it does not provide empirical data to test the reliability of the model in the local context of Indonesian higher education. In addition, there has been no in-depth exploration of the contextual impacts, such as academic culture, the experience of practicum facilitators, and diverse student characteristics, on the performance of these models.

In the context of higher education based on science and skills, such as in the Medical and Nursing Study Programs, the application of an accurate measurement model is very crucial, especially when associated with the integration of Islamic Religious Education (PAI) values (Beckett et al., 2017). Practicums in these two study programs not only assess cognitive aspects and clinical skills, but also emphasize affective aspects and professional spirituality, such as empathy, medical ethics, honesty, responsibility, and awareness of human and divine values (Matthews et al., 2023).

In this framework, the Graded Response Model (GRM) (Diebolt et al., 2023) has great potential to be applied in measuring aspects of students' Islamic attitudes and values during the practicum process, such as discipline in following schedules, consistency in maintaining ethical communication with patients, or politeness to colleagues and instructors (Ferrão et al., 2021)(Chen & Fujimoto, 2022). A Likert scale based on behavioral observation can be used to measure indicators of PAI values that are internalized into clinical practice, and the GRM can map the level of mastery progressively. For example, the score for attitudes toward patients in sensitive situations can be divided into "fairly ethical," "good," and "very good" in responding to patients' spiritual needs (Belous et al., 2021).

Meanwhile, GPCM is more suitable for use in performative assessments of clinical or laboratory tasks that have explicit Islamic value components. For example, in nursing procedures related to the care of corpses, examination of Muslim patients, or the application of health fiqh principles, the assessment rubric can include integrated cognitive, psychomotor, and Sharia values dimensions. Partial scores at each step of task implementation (e.g., readiness of equipment, intention before action, ethics in touching patients, and closing prayer) reflect Islamic practices in professional actions (Tutz et al., 2018). GPCM allows each aspect to be recognized separately and contribute to the total score proportionally (Leventhal, 2019).

Furthermore, the application of GRM and GPCM in practicum-based PAI assessments in the Medical and Nursing Study Programs encourages an integrative approach between science and spirituality (Marshall et al., 2020). This is in line with the goals of Islamic education, which not only

emphasize intellectual aspects but also character and personality development based on the values of tauhid. In practice, the assessment of spiritual communication skills, such as guiding patients in prayer, providing Islamic psychosocial support, or treating patients with compassion and manners, can be measured systematically with these models (Neal et al., 2023).

Thus, both GRM and GPCM can be used to support contextual, measurable, and competency-based assessment of Islamic religious education, in line with the direction of integrative curriculum transformation, which is currently the policy of many value-based higher education institutions. The use of this IRT model also strengthens the position of PAI as a relevant discipline in the development of value-based professional skills in the context of modern health services.

CONCLUSION

This study compares the Graded Response Model (GRM) and the Generalized Partial Credit Model (GPCM) in creating assessment tools for practical courses in higher education. The results show that each model has its own strengths. GPCM is better for tasks that give partial credit and don't follow a strict order, which is common in hands-on or performance-based assessments. GRM works better when the answers follow a clear order, such as different levels of skill or understanding. When estimating student ability, GPCM gives more stable results even if the data changes a little. GRM is more sensitive to differences between response levels, which helps detect small changes in student performance. GRM is useful when questions have clear levels of difficulty, while GPCM is more flexible, especially when scoring is subjective or when not all students reach the top score. Both models fit the data well. However, choosing the best model depends not only on statistics but also on the type of assessment, how it's scored, and what it's meant to measure. This study shows that there is no single best model for all situations. Instead, assessment should be flexible and fit the real learning context, especially in practical courses.

REFERENCES

Al-Taweel, D., Moreau, P., Koshy, S., Khedr, M. A., Nafee, N., Al-Romaiyan, A., Bayoud, T., Alghanem, S. S., Al-Awadhi, F. H., Al-Haqan, A., & Al-Owayesh, M. S. (2024). Empowering competence: A Program-Wide Active Learning Framework for a Pharmacy Program. American Journal of Pharmaceutical Education, 88(10). https://doi.org/10.1016/j.ajpe.2024.101272

Al Fariz, A. B. (2024). Mplus and the R MIRT Package: A Comparison of Model Parameter Estimation for Generalized Partial Credit Model (GPCM). Jurnal Pengukuran Psikologi dan Pendidikan Indonesia, 13(2), 162–179. https://doi.org/10.15408/jp3i.v13i2.40344

AlTfaili, H., Lamb, R. J., & Ginsburg, B. C. (2024). Assessment of Reduction in Stimulus

Generalization of Ethanol-Seeking During Recovery: A Rapid Procedure. Alcohol, 121, 161 – 167. https://doi.org/10.1016/j.alcohol.2024.09.003

Andersson, B. (2018). Asymptotic Variance of Linking Coefficient Estimators for Polytomous IRT Models. Applied Psychological Measurement, 42(3), 192 – 205. https://doi.org/10.1177/0146621617721249

Baylor, C., Bamer, A., Brown, C., Jin, J. L., Teixeira, J., & Nuara, M. (2024). The Communicative Participation Item Bank–Gender-Diverse Version: Item Bank Calibration and Short Form. American Journal of Speech-Language Pathology, 33(2), 952 – 968. https://doi.org/10.1044/2023_AJSLP-23-00260

Beckett, R. D., Etheridge, K., & DeLellis, T. (2017). A Team, Case-Based Examination and its Impact on Student Performance in a Patient Safety and Informatics Course. American Journal of Pharmaceutical Education, 81(6). https://doi.org/10.5688/ajpe816117

Belous, C. K., Wampler, R. S., & Ledford, B. L. (2021). Testing the Gay and Lesbian Relationship Satisfaction Scale with Item Response Modeling. Journal of Couple and Relationship Therapy, 20(1), 1 – 14. https://doi.org/10.1080/15332691.2020.1746460

Bisgaard, E., Clark, A., Hester, C., Napier, R., Grant, J., Scielzo, S., & Abdelfattah, K. (2021). Resident Engagement in a Wellness Program in a Large Academic Residency: A Follow-Up After Two Years of Wellness. Journal of Surgical Education, 78(5), 1430 – 1437. https://doi.org/10.1016/j.jsurg.2021.01.013

Bohnen, J. D., Demetri, L., Fuentes, E., Butler, K., Askari, R., Anand, R. J., Petrusa, E., Kaafarani, H. M. A., Yeh, D. D., Saillant, N., King, D., Briggs, S., Velmahos, G. C., & Moya, M. de. (2018). High-Fidelity Emergency Department Thoracotomy Simulator with Beating-Heart Technology and OSATS Tool Improves Trainee Confidence and Distinguishes Level of Skill. Journal of Surgical Education, 75(5), 1357 – 1366. https://doi.org/10.1016/j.jsurg.2018.02.001

Buchholz, J., & Hartig, J. (2019). Comparing Attitudes Across Groups: An IRT-Based Item-Fit Statistic for the Analysis of Measurement Invariance. Applied Psychological Measurement, 43(3), 241 – 250. https://doi.org/10.1177/0146621617748323

Bürkner, P. C., Schwabe, R., & Holling, H. (2019). Optimal Designs for the Generalized Partial Credit Model. British Journal of Mathematical and Statistical Psychology, 72(2), 271–293. https://doi.org/10.1111/bmsp.12148

Chan, P. G., Schaheen, L. W., Chan, E. G., Cook, C. C., Luketich, J. D., & D'Cunha, J. (2018). Technology-Enhanced Simulation Improves Trainee Readiness Transitioning to Cardiothoracic Training. Journal of Surgical Education, 75(5), 1395 – 1402. https://doi.org/10.1016/j.jsurg.2018.02.009

Chen, W., & Fujimoto, K. A. (2022). An Empirical Identification Issue of the Bifactor Item Response Theory Model. Applied Psychological Measurement, 46(8), 675 – 689. https://doi.org/10.1177/01466216221108133

Cummings, S. N., & Theodore, R. M. (2023). Hearing is Believing: Lexically Guided Perceptual Learning is Graded to Reflect the Quantity of Evidence in Speech Input. Cognition, 235. https://doi.org/10.1016/j.cognition.2023.105404

Dai, S., Vo, T. T., Kehinde, O. J., He, H., Xue, Y., Demir, C., & Wang, X. (2021). Performance of Polytomous IRT Models with Rating Scale Data: An Investigation Over Sample Size, Instrument Length, and Missing Data. Frontiers in Education, 6(September), 1–18. https://doi.org/10.3389/feduc.2021.721963

Diebolt, J. H., Cullom, M. E., Hornick, M. M., Francis, C. L., Villwock, J. A., & Berbel, G. (2023). Implementation of a Near-Peer Surgical Anatomy Teaching Program into the Surgery Clerkship. Journal of Surgical Education, 80(1), 1 – 6. https://doi.org/10.1016/j.jsurg.2022.08.005

Falk, C. F. (2020). The Monotonic Polynomial Graded Response Model: Implementation and a Comparative Study. Applied Psychological Measurement, 44(6), 465–481. https://doi.org/10.1177/0146621620909897

Ferrando, P. J., & Navarro-González, D. (2020). InDisc: An R Package for Assessing Person and Item Discrimination in Typical-Response Measures. Applied Psychological Measurement, 44(4), 327 – 328. https://doi.org/10.1177/0146621620909901

Ferrão, M. E., Bastos, A., & Alves, M. T. G. (2021). A Measure of Child Exposure to Household Material Deprivation: Empirical Evidence from the Portuguese Eu-Silc. Child Indicators Research, 14(1), 217 – 237. https://doi.org/10.1007/s12187-020-09754-4

Heriman, M., Dede Atung, Endang Sutisna, Nia Nurhayati, & Ika Kartika. (2024). Pengembangan Kurikulum Berbasis Keterampilan Abad ke-21: Perspektif dan Tantangan. Reslaj: Religion Education Social Laa Roiba Journal, 6(6), 2724–2741. https://doi.org/10.47467/reslaj.v6i6.1709

Hermita, N., Putra, Z. H., Alim, J. A., Wijaya, T. T., Anggoro, S., & Diniya, D. (2021). Elementary Teachers' Perceptions on Genially Learning Media Using Item Response Theory (IRT). Indonesian Journal on Learning and Advanced Education (IJOLAE), 4(1), 1–20. https://doi.org/10.23917/ijolae.v4i1.14757

Huang, J., Shu, T., Dong, Y., & Zhu, D. (2023). Constructing and Validating a Self-Assessment Scale for Chinese College English-Major Students' Feedback Knowledge Repertoire in EFL Academic Writing: Item Response Theory and Factor Analysis Approaches. Assessing Writing, 56(March), 100716. https://doi.org/10.1016/j.asw.2023.100716

Huggins, S. (2017). Practice-Based Learning in Higher Education. Library Trends, 66(1), 1–12. https://doi.org/10.1353/lib.2017.0024

Idiyatova, Y., Sadvakassova, A., & Mukhatayev, A. (2024). Modern Approaches to the Assessment of Academic Achievements in Higher Education Institutions: a Systematic Review of the Literature. Deleted Journal, 46(2), 25–37. https://doi.org/10.59787/2413-5488-2024-46-2-25-37

Jonas, K. G., & Markon, K. E. (2019). Modeling Response Style Using Vignettes and Person-Specific Item Response Theory. Applied Psychological Measurement, 43(1), 3 – 17. https://doi.org/10.1177/0146621618798663

Joo, S.-H., Lee, P., & Stark, S. (2022). Bayesian Approaches for Detecting Differential Item Functioning Using the Generalized Graded Unfolding Model. Applied Psychological Measurement, 46(2), 98 – 115. https://doi.org/10.1177/01466216211066606

Kirkup, M. L., Adams, B. N., Meadows, M. L., & Jackson, R. (2016). Development and Implementation of an Electronic Clinical Formative Assessment: Dental faculty and student perspectives. Journal of Dental Education, 80(6), 652 – 661. https://www.scopus.com/inward/record.uri?eid=2-s2.0-84973303642&partnerID=40&md5=77c9155bfa34b2c5833710522df991a5

Kulkarni, M. M., Kamath, V. G., Kamath, A., Lewis, S., Bogdanovica, I., Bains, M., Cranwell, J., Fogarty, A., Arora, M., Nazar, G. P., Ballal, K., Bhagwath, R., & Britton, J. (2021). Exposure to Tobacco Imagery in Popular Films and the Risk of Ever Smoking Among Children in Southern India. Tobacco Control, 30(5), 560 – 566. https://doi.org/10.1136/tobaccocontrol-2019-055353

Kurnia, A. (2019). Analisis Tes Kemampuan Berpikir Kritis Matematis Siswa dengan Menggunakan Generalized Partial Credit Model (GPCM): Penelitian Deskriptif Kuantitatif di SMP …. Pediamatika: Journal of Mathematical Science and Mathematics Education, 01(02), 105–114. http://digilib.uinsgd.ac.id/22038/

Larsson, J., Dencker, M., Bremander, A., & Olsson, M. C. (2022). Cardiorespiratory Responses of load Carriage in Female and Male Soldiers. Applied Ergonomics, 101. https://doi.org/10.1016/j.apergo.2022.103710

Learning, M., Ai, E., Prediction, D., Review, S., Ndlovu, B., Maguraushe, K., & Mabikwa, O. (2025). The Indonesian Journal of Computer Science. 14(2), 2357–2386.

Leventhal, B. C. (2019). Extreme Response Style: A Simulation Study Comparison of Three Multidimensional Item Response Models. Applied Psychological Measurement, 43(4), 322 – 335. https://doi.org/10.1177/0146621618789392

Liu, S. H., Chen, Y., Bellinger, D., de Water, E., Horton, M., Téllez-Rojo, M. M., & Wright, R. (2024). Pre-natal and Early Life Lead Exposure and Childhood Inhibitory Control: an Item Response Theory Approach to Improve Measurement Precision of Inhibitory Control. Environmental Health: A Global Access Science Source, 23(1), 1–13. https://doi.org/10.1186/s12940-023-01015-5

Liu, Z., Li, Y., & Wang, J. (2025). Exploring the Flexibility of Word Position Encoding in Chinese Reading: the Role of Transposition Effects. Language, Cognition and Neuroscience, 40(2), 263 – 269. https://doi.org/10.1080/23273798.2024.2417430

Lozoya-Santos, J. D. J., Guajardo-Leal, B. E., Vargas-Martínez, A., Molina-Gaytán, I. E., Román-Flores, A., Ramirez-Mendoza, R., & Morales-Menendez, R. (2019). Knowledge Generation in Higher Education Institutions. IEEE Global Engineering Education Conference, Educon, April 2019, 628–633. https://doi.org/10.1109/Educon.2019.8725273

Lubbe, D., & Schuster, C. (2019). A Graded Response Model Framework for Questionnaires with Uniform Response Formats. Applied Psychological Measurement, 43(4), 290 – 302. https://doi.org/10.1177/0146621618789394

Lubbe, D., & Schuster, C. (2020). A Scaled Threshold Model for Measuring Extreme Response Style. Journal of Educational and Behavioral Statistics, 45(1), 86 – 107. https://doi.org/10.3102/1076998619859541

Luu, N. N., Yver, C. M., Douglas, J. E., Tasche, K. K., Thakkar, P. G., & Rajasekaran, K. (2021). Assessment of YouTube as an Educational Tool in Teaching Key Indicator Cases in Otolaryngology During the COVID-19 Pandemic and Beyond: Neck Dissection. Journal of Surgical Education, 78(1), 214 – 231. https://doi.org/10.1016/j.jsurg.2020.06.019

Marshall, L. L., Kinsey, J., Nykamp, D., & Momary, K. (2020). Evaluating Practice Readiness of Advanced Pharmacy Practice Experience Students Using the Core Entrustable Professional Activities. American Journal of Pharmaceutical Education, 84(10), 1292 – 1299. https://doi.org/10.5688/ajpe7853

Matthews, D. E., Kelley, K. A., Li, J., & Beatty, S. (2023). Improving Knowledge of Top 200 Medications Through Retrieval Practice, Content Alignment, and Autonomous Learning. American Journal of Pharmaceutical Education, 87(3), 356 – 363. https://doi.org/10.5688/ajpe9079

Mehnert, J. M., Silk, A. W., Lee, J. H., Dudek, L., Jeong, B. S., Li, J., Schenkel, J. M., Sadimin, E., Kane, M., Lin, H., Shih, W. J., Zloza, A., Chen, S., & Goydos, J. S. (2018). A phase II trial of riluzole, an antagonist of metabotropic glutamate receptor 1 (GRM1) signaling, in patients with advanced melanoma. Pigment Cell and Melanoma Research, 31(4), 534–540. https://doi.org/10.1111/pcmr.12694

Mirunnisa, M., & Razi, Z. (2021). Analysis of High School Students 'Mathematic Critical Thinking Ability with Graded Response Models. Budapest International Research and Critics Institute (BIRCI-Journal): Humanities and Social Sciences, 4(1), 1108–1116. https://doi.org/10.33258/birci.v4i1.1719

Model, I., & Evaluation, D. (2024). Langgam: Instrument Model Development Evaluation Authentic. 3(2), 8–13.

Naveiras, M., & Cho, S.-J. (2023). Using Auxiliary Item Information in the Item Parameter Estimation of a Graded Response Model for a Small to Medium Sample Size: Empirical Versus Hierarchical

Bayes Estimation. Applied Psychological Measurement, 47(7–8), 478 – 495. https://doi.org/10.1177/01466216231209758

Neal, C. J., Durning, S. J., Dharmapurikar, R., McDaniel, K. E., Lad, S. P., & Haglund, M. M. (2023). From Their Eyes: What Constitutes Quality Formative Written Feedback for Neurosurgery Residents. Journal of Surgical Education, 80(3), 323 – 330. https://doi.org/10.1016/j.jsurg.2022.10.003

Phillips, D. B., Ehnes, C. M., Welch, B. G., Lee, L. N., Simin, I., & Petersen, S. R. (2018). Influence of Work Clothing on Physiological Responses and Performance During Treadmill Exercise and the Wildland Firefighter Pack Test. Applied Ergonomics, 68, 313 – 318. https://doi.org/10.1016/j.apergo.2017.12.010

Qian, Z., Yang, Y., Tan, J., Li, Y., Zhou, J., Huang, J., Assanangkornchai, S., & Chen, J. (2025). Psychometric Evaluation of the Chinese Version of the Mental Health System Responsiveness Questionnaire for Psychiatric Outpatients: Classical Test Theory and Item Response Theory Approaches. Patient Preference and Adherence, 19, 729 – 740. https://doi.org/10.2147/PPA.S503016

Reimers, J., Turner, R. C., Tendeiro, J. N., Lo, W.-J., & Keiffer, E. (2023). The Effects of Aberrant Responding on Model-Fit Assuming Different Underlying Response Processes. Applied Psychological Measurement, 47(5–6), 420 – 437. https://doi.org/10.1177/01466216231201987

Ren, D. M., Abrams, A., Banigan, M., Batabyal, R., Chamberlain, J. M., Garrow, A., Izem, R., Nicholson, L., Ottolini, M., Patterson, M., Sarnacki, R., Walsh, H. A., & Zaveri, P. (2022). Evaluation of Communication and Safety Behaviors During Hospital-Wide Code Response Simulation. Simulation in Healthcare, 17(1), E45–E50. https://doi.org/10.1097/SIH.0000000000000575

Sahu, S.K., Bass, M. R., Sabariego, C., Cieza, A., Fellinghauer, C. S., & Chatterji, S. (2020). A full Bayesian Implementation of a Generalized Partial Credit Model with an Application to an international disability survey. Journal of the Royal Statistical Society. Series C: Applied Statistics, 69(1), 131–150. https://doi.org/10.1111/rssc.12385

Santos, J. S., Andrade, W. L., Brunet, J., & Araujo Melo, M. R. (2020). A Systematic Literature Review of Methodology of Learning Evaluation Based on Item Response Theory in the Context of Programming Teaching. Proceedings - Frontiers in Education Conference, FIE, 2020-Octob. https://doi.org/10.1109/FIE44824.2020.9274068

Sultan, S., & Zhang, Z. (2023). A Stable Time-Dependent Mesh Method for Generalized Credit Rating Migration Problem. Journal of Nonlinear Mathematical Physics, 30(4), 1774–1803. https://doi.org/10.1007/s44198-023-00157-x

Tendeiro, J. N., & Castro-Alvarez, S. (2019). GGUM: An R Package for Fitting the Generalized Graded Unfolding Model. Applied Psychological Measurement, 43(2), 172 – 173. https://doi.org/10.1177/0146621618772290

Thissen, D. (2015). Psychometrics: Item Response Theory. In International Encyclopedia of the Social & Behavioral Sciences: Second Edition (Second Edi, Vol. 19). Elsevier. https://doi.org/10.1016/B978-0-08-097086-8.42071-4

Tu, N., Zhang, B., Angrave, L., & Sun, T. (2021). bmggum: An R Package for Bayesian Estimation of the Multidimensional Generalized Graded Unfolding Model with Covariates. Applied Psychological Measurement, 45(7–8), 553 – 555. https://doi.org/10.1177/01466216211040488

Tutz, G., Schauberger, G., & Berger, M. (2018). Response Styles in the Partial Credit Model. Applied Psychological Measurement, 42(6), 407 – 427. https://doi.org/10.1177/0146621617748322

Umucu, E., Brooks, J. M., Lee, B., Iwanaga, K., Wu, J.-R., Chen, A., & Chan, F. (2018). Measuring dispositional optimism in student Veterans: An item response theory analysis. Military

Psychology, 30(6), 590 – 597. https://doi.org/10.1080/08995605.2018.1522161

Von Davier, M., & Yamamoto, K. (2004). Partially observed mixtures of IRT models: An extension of the generalized partial-credit model. Applied Psychological Measurement, 28(6), 389–406. https://doi.org/10.1177/0146621604268734

Wallmark, J., Ramsay, J. O., Li, J., & Wiberg, M. (2023). Analyzing Polytomous Test Data: A Comparison Between an Information-Based IRT Model and the Generalized Partial Credit Model. Journal of Educational and Behavioral Statistics, 753–779. https://doi.org/10.3102/10769986231207879

Wallmark, J., Ramsay, J. O., Li, J., & Wiberg, M. (2024). Analyzing Polytomous Test Data: A Comparison Between an Information-Based IRT Model and the Generalized Partial Credit Model. Journal of Educational and Behavioral Statistics, 49(5), 753 – 779. https://doi.org/10.3102/10769986231207879

Wei, J., Cai, Y., & Tu, D. (2023). A Mixed Sequential IRT Model for Mixed-Format Items. Applied Psychological Measurement, 47(4), 259 – 274. https://doi.org/10.1177/01466216231165302

Wijayanto, F., Bucur, I. G., Groot, P., & Heskes, T. (2023). autoRasch: An R Package to Do Semi-Automated Rasch Analysis. Applied Psychological Measurement, 47(1), 83 – 85. https://doi.org/10.1177/01466216221125178

Winiger, S., Singmann, H., & Kellen, D. (2021). Bias in Confidence: A Critical Test for Discrete-State Models of Change Detection. Journal of Experimental Psychology: Learning, Memory, and Cognition, 47(3), 387 – 401. https://doi.org/10.1037/xlm0000959

Zhang, Z. (2021). Asymptotic Standard Errors of Generalized Partial Credit Model True Score Equating Using Characteristic Curve Methods. Applied Psychological Measurement, 45(5), 331 – 345. https://doi.org/10.1177/01466216211013101